NEXT-GEN SEARCH ENGINES

By AMARDEEP GUPTA
HEAD, DEPTT. OF COMPUTER SCIENCE, D A V COLLEGE, AMRITSAR

Current search engines--even the constantly surprising Google--seem unable to leap the next big barrier in search: the trillions of bytes of dynamically generated data created by individual Web sites around the world, or what some researchers call the "deep web."

The challenge that we have right now is not information overload, it's information overlook.

The latest interface enhancement from the major portals is the introduction of tabs, long a Google staple. Yahoo! recently introduced a longer search bar on its front page and added tabs for images, the yellow pages, and products. AOL followed suit with a new "In Your Area" search, which searches AOL's Yellow Pages. Google's tabs are the most numerous of the bunch, featuring Web, images, groups, directory, and news.

Will any of these ideas show up on the search engines of the future?

Imagine a world where you type something into a search engine and the result is a uniquely indexed, real-time display of links generated just for you.

Currently, the search engine market share breaks out among the top four players in the following way: Google-32%; Yahoo-26%; AOL-19%; MSN-17%. However, Google's results are currently displayed not only on Google, but also on Yahoo and AOL resulting in a 76% market share in the search space for Google. But this is about to change. Yahoo has purchased Inktomi and is on the brink of ditching the Google search results in favor of Inktomi's results splitting the market more evenly. And while no one is sure what Microsoft is up to with MSN search, the experts seem to agree that the company will come out with a competitive search technology at some point in the not-too-distant future.

Chronology of Search Engines

The first generation of Web search tools used on-the-page relevancy ranking, creating algorithms based on location and frequency of keywords. First generation added relevancy for META tags, keywords in the domain name, and a few bonus points for having keywords in the URL. Basic spam filters emerged that got rid of keyword stuffing and same color text. The portals also made their appearance, and engines started looking like giant billboards and overstuffed yellow pages.

But, using keywords in various tags didn't help as much?

Instead, the engines took it a step further in their quest for relevant results by bringing in 2nd generation engines. Second generation, which is in full swing with the themes thing, added off-the-page relevancy, using hyperlinks and visit duration data for results ranking. A few of the major components they employ are tracking clicks, page reputation, link popularity, temporal tracking, and link quality. Then they started adding in term vectors, stats analysis, cache data, and context where two-word keyword pairs were extracted from a page to better categorize it.

The current state of search engines can be compared to a phone book, which is updated irregularly, and has most of the pages ripped out.

- The engines are limited by network bandwidth, disk storage, computational power, or a combination of these items.
- Lower bound on the size of the indexable Web is 320 million pages approximately.
- The coverage of the major Web search engines varies dramatically, but the coverage can be increased by combining the results of multiple engines.
- The indexing patterns of the engines vary and engines index only a fraction of the total number of documents on the Web.

SEARCH ENGINE INDUSTRY IN FORTHCOMING YEARS

The year 2003 was, generally speaking, a good year for the search engine industry. 2004 will be even better, but with

some major changes, like upcoming of a few newcomers. 2004 will be a year where you can expect news to hit this industry on a daily basis. In 2003, hardly a day went by that there wasn't something happening at one company or another. Expect 2004 to be even busier.

Additionally, 2004 will be a year where the stakes are getting higher- much higher, and not just for Google! The level of competition among search engines will get tougher, in a race to land what is called "targeted eyeballs".

3G Search Engines

Third generation is already underway. It adds word stemming and a thesaurus on top of the term vector database to assist in keeping a search in context. Auto extraction of keyword pairs also helps automatically categorize a page, where searches like `shop for' or `find' trigger totally different search results based on the context or intent of the person doing the searching. G3 adds Web maps which, although not searchable, are a useful filtering tool to get rid of duplicate sites and many stand alone pages that drive traffic to only a few destinations.

They will also be extracting as much data as possible about your individual searching habits. All the major engines plan on building personal profiles, little robots that `come to know you' over a period of time, based on past searching habits.

◉ *What are "Theme" Engines?*

It's just another way of saying they are implementing `second generation' search engine strategies. Using a term vector database, they weigh page keyword density to calculate the page vector, which is compared and stored relative to the term vector. They then compute a Web page reputation by graphing interconnectivity and link relevancy, making sure the reputation of the page and the content on the page actually match. The closest matches get the highest search engine positioning.

Today all search engines are moving toward being theme-based.

How can we create Web pages that theme engines will like and boost our odds at getting top rankings?

- Gone are the days when a single Web site was used. Set up additional Web sites for different areas of your company, Interlink them together, carefully controlling how you're describing the links pointing to those other pages. Change the content and the featured theme, with same overall design, for the benefit of visitors to your site.
- When using link text, try eliminating punctuation marks and small, inconsequential words, like "and," "the," "it," and "for."
- Keep each page focused on one topic, and keep each site focused on one topic.
- "Pull all the pages out of the database, set them up as static pages, and put them to work for you in the search engines.
- Check each page carefully. Make sure that everything on the page points to one central theme or has one focus. Do everything you can to make sure that the engine understands what your primary theme is.
- Create your pages as if you're writing an article on your keyword phrase.
- META tags certainly don't hold the importance that they once did; so don't depend solely on them to achieve a top ranking.

Go after keyword phrases that aren't as competitive in the beginning. Then, go after the more competitive phrases next.

◉ Verticals and Content Engines: **The Bloated Engine?**

The search engines are indexing 20% of the available pages on the net, or 25%, or 30%… The search engines should be indexing fewer pages, not more.

Some possibilities that should be considered for universal finding tools going wrong:

1. Collection Policy: Search engines index anything and everything submitted to them, the everything's accepted approach has created unnecessary bulk to the search collections.
2. The lack of Human Indexing: Technology is great, but unless the general population is willing to take the time to master "Boolean Operators", the best methodology available is going to be plain old human classification.

3. No focus on Subject: Search engines are collecting on all subjects for all people increasing searcher's frustration.

○ *The Vertical Engine*

The evolution has now continued with the addition of Subject Specific Collections or Verticals.

These sites have taken the emphasis off of being all things to all people, and placed the focus on collecting for a single subject, and in some cases, a specific audience. Another important addition for the Verticals is the use of "Editors" and "Guides." Editors represent the first foray into the area of subject expertise. Now, there is an obvious critique regarding the use of untrained editors, but the combination of expertise with "human indexing" represents a marked improvement in the overall collection technique.

○ *Where Do We Go From Here?*

The searchable collection of the future will inevitably have more requirements for content inclusion. Many of the coming trends will likely mirror the theory of "Library Science" and how paper based collections are developed today. Some of the possibilities may include:

1. Search Sites will actively develop their collections rather than passively waiting for content submissions.
2. A formal Collection Policy will be posted and adhered to in defining submission and selection criteria. Content can still be used as a marketing tool, but the free-for-all has to stop.
3. Picking an audience and sticking with that audience will enable Web Sites to narrow the amount and quality of available content. A properly profiled audience may also be the marketable site of the future.
4. Sites may choose to collect content that meets a certain "structural format" such as short stories of more than 1500 words, or "subject format".

◉ **THE SEMANTIC WEB:** OPPORTUNITIES AND CHALLENGES FOR NEXT-GENERATION WEB APPLICATIONS

Recently there has been a growing interest in the investigation and development of the next generation web - the Semantic Web.

With the advent of the Semantic Web, resources on the Web will be represented semantically in ontologies. Semantics-based web search engines can be built in which each query is executed within the context of some ontology. The guidance from ontologies will increase recall and precision of the search result. For example, one might pose a query "return all the reviewers for book 'The Semantic Web: an Introduction'" to a semantics-based web search engine, then the engine will return only reviewers for this book instead of returning web pages that contain keyword "reviewer". It is worth mentioning that some systems that use ontologies to enhance web search engines have been developed. Since ontologies are built on a domain basis, web search engines might be also built on a domain basis, and hence metasearch engines, which interface with multiple remote search engines and select and rank remote search engines intelligently, might be very useful

What Does the Future Hold?

In the future, you might be able to load the engine full of lists of keywords. Your interests, likes and dislikes, geographical info, and favorite Web sites can be entered, from which the engine can create a context engine just for you. Just think, they'll know what your next search is likely to be, even before you do.

The future of searching will not only be about text, but will increasingly rely on visual models to help users understand the distribution of meaning and relationships between information sources.

Perhaps the most promising visual meta-search engine for educators is Kartoo (www.kartoo.com).

Kartoo is one of the most student-friendly and stable members of the new stable of visual search engines. If you are attached to Google you may want to check out the TouchGraph and Anacubis visual browsers for Google, as well as the Google Set Vista for visualizing Google sets. Instructional applications of the Google browsers are not as self-evident as with Kartoo, but advanced searchers should enjoy using the tools to play with their favorite searches. If you have money in your budget then you might be interested in the comprehensive (and visually stunning) Grokker, currently available as a preview release..

*Several Web search engines now offer an "**Image Search**" option.*

How image search engines work:

It's no coincidence that many of the image search engines are offered by the same services that also offer text indexing of the Web. The Web crawlers employed by AltaVista, Google, Lycos, etc. travel from Web site to Web site, pulling in the contents of Web pages. These pages form the basis for the familiar text searching indexes. Since image files linked to by the Web pages can be identified by MIME type or file extension (e.g. GIF, JPG or PNG) and downloaded by the same Web crawlers, cataloging images is a natural extension of what the search engines are already doing.

The difficulty comes in deciding how to index an image so it can be searched using text. The simplest and most automated way is to use the text "near" the image.

If all the image search engines based their indexing on the same text, one wouldn't expect much variation among them, but refinements to the process do make a difference. Although most of the sites don't go into much detail about how they create their indexes, those that do provide some insight into the process.

For example, Google claims that it "analyzes the text on the page adjacent to the image, the image caption and dozens of other factors to determine the image content" and that it "uses sophisticated algorithms to remove duplicates and ensure that the highest quality images are presented first in your results." In other words, steps are taken to try to improve the relevance of images displayed. Techniques might include giving heavier weighting to text more tightly bound to the image (such as the ALT tag) as opposed to text that simply appears on the same page.

Ditto, which is exclusively an image search engine, claims that it achieves improved relevance by employing "a proprietary filtering process that combines sophisticated automated filtering with human editors." Similarly, Picsearch claims that it has a "relevancy unrivalled on the web due to its patent-pending indexing algorithms." Other approaches to searching for images are available.

A few of the more mature technologies are available commercially, such as IBM's **Query by Image Content** and **Excalibur's Visual Retrieval Ware.**